## The nature of the forbidden/allow asymmetry: two correlational studies

Holleman, B.C.

***published in***
Sociological Methods and Research
1999

***DOI (link to publisher)***
10.1177/0049124199028002004

***document version***
Publisher's PDF, also known as Version of record

# The Nature of the Forbid/Allow Asymmetry : Two Correlational Studies

BREGJE HOLLEMAN

*Sociological Methods & Research* 1999 28: 209
DOI: 10.1177/0049124199028002004

*As the verbs* forbid *and* allow *are considered each other's counterparts, one would expect the answers to questions worded with forbid or allow to be each other's opposites. Research shows that this is not the case. Use of these verbs in surveys causes a wording effect known as the* forbid/allow asymmetry. *Findings do not reveal, however, where the asymmetry originates: during the stage of attitude localization or during the mapping of the attitude onto one of the response options. In this article, a correlational design was used. Two experiments were carried out focusing on the congenericity of forbid and allow, one on attitudes toward environmental issues and a replication on attitudes concerning ethnic groups. Results of both experiments show that forbid and allow questions are congeneric; that is, they measure the same attitude. Answers to forbid/allow questions reflect similar attitudes that are expressed differently on the answering scales due to the use of both verbs. In addition, the explanation of the effect focusing on the answering behavior of indifferent respondents is discussed and explored.*

# The Nature of the Forbid/Allow Asymmetry

## Two Correlational Studies

BREGJE HOLLEMAN
*Free University Amsterdam*

Research has repeatedly demonstrated that small changes in the wording of a question cause huge differences in the responses obtained (Schuman and Presser 1981; Molenaar 1982; Jobe and Mingay 1991). This raises questions about the validity of survey questions: Which particular question wording measures what the questionnaire designer intends to measure? The basic goal of research on wording effects is traditionally to generate practical advice for questionnaire design (Billiet 1989), but equally important is the more fundamental goal to understand the cognitive processes underlying question answering and the variables that affect responses (Cicourel 1982; Jobe and Mingay 1991). Theoretical insight is a prerequisite for providing solid practical advice. We can only advise on question

AUTHOR'S NOTE: *I am very grateful to Huub van den Bergh for his guidance and help with the statistics. I also wish to thank Jaak Billiet, Herb Clark, Paul van den Hoven, Willem Saris, Arie Verhagen, and the anonymous reviewers for their elaborate and very helpful comments on a previous draft of this article.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 28 No. 2, November 1999 209-244

**TABLE 1: Forbid/Allow Experiment Reported by Rugg (1941)**

| | |
|---|---|
| Do you think the United States should forbid public speeches against democracy? | |
| Yes, forbid | 54% |
| No, not forbid | 46% |
| Do you think the United States should allow public speeches against democracy? | |
| Yes, allow | 25% |
| No, not allow | 75% |

wording if we know how wording variation causes differences in responses and in what respect exactly two seemingly equivalent questions measure something different.

A wording effect that has received much attention for more than half a century of research is the *forbid/allow asymmetry,* identified by Rugg (1941). Rugg found that respondents were more likely to support freedom of speech when the question was worded with the verb forbid, resulting in a difference of 21 percent between answers to two questions that seem logically equivalent (see Table 1).

Since Rugg's finding, the effect has been replicated in the United States (Schuman and Presser 1981; Bishop et al. 1988), Germany (Bishop et al. 1988; Hippler and Schwarz 1986), and Belgium (Waterplas, Billiet, and Loosveldt 1988) over a wide range of issues using a variety of administration modes (phone, face-to-face or self-administered questionnaires). Very often a significant difference between responses to forbid/allow questions is found. Schuman and Presser (1981:296) note that this wording effect caused the greatest difference in responses, up to 30 percentage points. The results of these replications, however, are not equivocal. The effect does not occur in every experiment, nor is it necessarily in the expected direction. Close analysis of the experiments shows that the replications are less systematic than they seem to be. For example, in almost every experiment, different question contents or administration modes are used (Waterplas et al. 1988).[1]

Many explanations for the forbid/allow asymmetry have been suggested, two of which will be summarized briefly. One explanation focuses on the extreme connotations of both forbid and allow: "The former sounds harsher and may therefore be more difficult to endorse, whereas the latter in some context might seem to encourage a deviant

behavior and therefore may invite opposition" (Schuman and Presser 1981:296); in other words, "To answer 'no' to forbid does not mean the same as 'yes' to allow: to agree to forbid implies a real act of opposition, but to disagree to allow means merely to abstain from support" (Clark and Schober 1992:31). This hypothesis was supported by Hippler and Schwarz (1986), who asked respondents to rate the extremeness of the fictitious opinion that "Mr. X should forbid (not allow, allow, not forbid) peep shows" on a bipolar scale. Allowing and forbidding were judged to be more extreme than not forbidding and not allowing, respectively. Differences were small, however, and there was no doubt that not forbidding is considered a statement in favor of the issue and that not allowing is considered a statement in opposition of it (Hippler and Schwarz 1986:91).

The second hypothesis introduced by Hippler and Schwarz (1986) elaborates on the connotations hypothesis by focusing on the role of *indifferent respondents*. Because of the extreme connotations of forbid and allow, respondents holding a weak attitude may be particularly unlikely to endorse either the forbid or the allow statement, responding "no" to both question forms. Social cognition research shows that individuals pay more attention to positive instances of a behavior than to the absence of a behavior (Nisbett and Ross 1980). This "tendency of . . . humans to exhibit greater difficulty in the processing of nonoccurrences than occurrences" (Fazio, Sherman, and Herr 1982:404) is referred to as the *feature-positive effect*. "Similarly, when asked if something should be done [allowed or forbidden], individuals may focus on the implications of the behavior under consideration, and they may not consider the implications of the absence of this behavior . . . thus resulting in the feeling that one has only said it should not be allowed, without ever commenting on whether it should be forbidden" (Hippler and Schwarz 1986:89). This explanation predicts that the forbid/allow asymmetry is primarily due to indifferent or ambivalent respondents and should not be obtained for respondents holding a clear pro or con attitude. Results of experiments by Hippler and Schwarz (1986) and Waterplas et al. (1988) are consistent with this explanation: When subjects holding a weak or indifferent attitude toward the issue were removed from the sample, the wording effect disappeared.

*THEORETICAL PROBLEMS AND SOLUTIONS*

Both hypotheses are promising but suffer from some vagueness. The connotations hypothesis does not provide any real insight into the cognitive mechanisms underlying the asymmetry, causing it to remain a post hoc explanation rather than a testable hypothesis. However, the connotations hypothesis can be transformed into a testable hypothesis by relating it to more general theories of the question/answer process.

Generally, four stages are distinguished within the question/answer process: interpretation of the question, location of the relevant attitude structure, retrieval of the attitude (or formation of the attitude), and fitting of the judgment onto one of the precoded answering categories (cf. Tourangeau and Rasinski 1988; Sudman, Bradburn, and Schwarz 1996). The connotations hypothesis does not make explicit whether the difference between forbid/allow answers is caused in the first two stages of the question/answer process (during attitude localization and retrieval, or formation), or whether similar attitudes are being measured and the asymmetry stems from the last stage, in which the opinion is mapped onto the answering options "yes" or "no." Does the explanation, as worded by Schuman and Presser (1981), mean that answers to forbid questions reflect different attitudes than answers to equivalent allow questions? The connotations, or semantic fields of both verbs in general, may be so strong that not only the attitude toward a specific issue (e.g., abortion) is measured but also a general attitude toward forbidding or allowing. This could be called a *different attitudes hypothesis*. But it may also be the case that the asymmetry results from slight changes in perceptions of the meanings of attitude question response options. Krosnick and Schuman (1988:940) theorize that the asymmetry is caused by differences in the way respondents map their answers to response options due to the use of forbid and allow: "'Not allowing' is perceived as a less extreme stance than is 'forbidding.'" This could be called a *different scales hypothesis*.

This distinction between different attitudes being measured versus similar attitudes being expressed differently in response options is relevant not only for a better understanding of the question/answer process but for practice as well. If forbid/allow questions measure different attitudes, they differ in validity. Either, or both, may not measure what the researcher intends to measure. But if forbid/allow questions

measure a similar attitude, then they are equally valid, even though the answers to the questions differ. In that case, the questionnaire designer's problem is how to translate the yes/no answers back correctly to the true attitude.

The theoretical vagueness of the connotations hypothesis causes the *indifferent respondents hypothesis* to be vague as well, since the latter is based on the former. First, the correct version of connotation hypothesis must be found (different attitudes or different scales). As will be discussed in the next section, psychometric criteria can be developed to test both hypotheses. Only when insight into the exact nature of the differences between forbid/allow answers is established can the next question be posed: Exactly how do the answers to forbid and allow questions differ for the subgroup of indifferent respondents?

### *MEASUREMENT PROBLEMS AND SOLUTIONS*

To gain precise insight into the nature and cause of the forbid/allow asymmetry, a research design has to allow for detection of the exact differences between responses. The general aim of wording effect research is to gain an understanding of the relation between the answers respondents give and the respondents' true attitude. The main issue is whether questions worded differently, although conceptually equivalent, measure similar attitudes (a test of the different attitudes hypothesis) and, if so, whether the similar attitudes are expressed on similar scales (the different scales hypothesis). Furthermore, it is important to use a research design that enables one to generalize beyond the level of the specific questions that were used in a particular experiment. As will be argued in this section, the research design that is generally used in forbid/allow research does not comply with these two demands.

In Rugg's (1941) experiment, and in all other experiments on the forbid/allow asymmetry, a split ballot design is used analyzing each question separately. Two random halves of a population sample each get a variant of the same question. If a significant difference between responses is found, it is concluded that the wording of the question has a systematic effect on the response (Molenaar 1986).

The first problem of this research design is that although it can determine whether there is a difference somewhere in the question/answer process, it does not allow one to detect whether differences in forbid/allow answers reflect differences in the attitudes measured, or differences in the answering scales due to the use of forbid or allow. By just comparing the observed scores, as is done in traditional split ballot designs, two assumptions are being made that are not necessarily true. First, some authors assume that the answers to both questions are represented on identical scales, so that differences in observed scores are interpreted as differences in true scores and as differences in attitudes. Others assume that the answers reflect similar attitudes, so that differences in observed scores are interpreted as differences in the meaning of the response options. Both assumptions should be tested first. Second, when comparing observed scores, it is assumed that reliabilities (and error score components) of both questions are the same. This is also not necessarily the case. The question of whether similar attitudes are being measured can be answered only if the true scores to forbid/allow questions are compared. But true score components and error score components cannot be distinguished when analyzing each question, which was measured at only one time, separately, which makes a traditional split ballot inappropriate.

A second problem with split ballot designs when analyzing single questions separately is generalization. By computing differences between responses to variants of single questions, it is impossible to generalize beyond the level of the specific linguistic variables that were used in an experiment (Molenaar 1986), thus causing a generalization problem similar to the language-as-fixed-effect fallacy (Clark 1973). The chosen operationalization of possible question contents, wording variation, administration mode, and so on has to be a good sample of possible variations in forbid/allow contexts in order to be able to generalize beyond specific contexts. Such an operationalization is never possible when experiments consist of only one manipulated question. Comparing the answers to a cluster of forbid questions with the answers to a cluster of allow questions would solve this generalization problem to some extent.

A design in which error scores and true scores can be distinguished is a correlational design. The question concerning the difference in attitudes measured is an issue of *congenericity* (Jöreskog 1971). Two

tests, or items, are congeneric if they measure the same trait, irrespective of errors of measurement. In other words, two items are decided to be congeneric if the correlation coefficient between the true scores of those items is unity (equals 1.0).

The different attitudes hypothesis will be confirmed, then, if forbid and allow questions turn out to be noncongeneric. Conversely, if forbid and allow questions correlate unity, they measure similar attitudes. If, despite that, the observed scores for forbid and allow questions differ, then the different scales hypothesis is confirmed: Because the two verbs carry extreme connotations, an equivalent opinion is expressed differently on the answering scale. Due to the wording of the question, the response scales for forbid and allow are dissimilar interval scales. The response scales can differ in two possible ways: Either the midpoints of the scales can differ or the midpoints of the scales and the distances between the intervals can differ. If only the midpoints of the interval scales differ, the error score variances and true score variances will be the same and the observed scores for forbid and allow will be different.[2] If the size of the intervals differs as well as the midpoints, the observed scores and the true score variances for forbid and allow will differ. In both cases, forbid and allow will still be congeneric.

The explanation based on the feature-positive effect argues that the asymmetry is primarily caused by the responding behavior of indifferent respondents, since indifferents are hypothesized not to consider the fact that forbidding implies a consistent attitude toward allowing. This indifferent respondents hypothesis could predict that forbid and allow are noncongeneric for the subsample of indifferents while being congeneric for the other respondents.

### *TOWARD AN EXPERIMENTAL CORRELATIONAL DESIGN*

The most important hypothesis in this research is the different attitudes hypothesis, focusing on the congenericity of forbid/allow questions. The other hypotheses gain importance once it turns out that forbid/allow questions measure the same trait. Only then are differences in reliabilities or scales between forbid and allow questions meaningful.

To answer this question of congenericity, a correlational measure between forbid and allow has to be obtained that is corrected for

random error. This can be achieved through the concept of construct validity, which requires the construction of a questionnaire consisting of several questions. If a cluster of questions measuring one underlying construct (e.g., attitude toward abortion) were constructed in forbid and allow versions, respectively, one could compute reliabilities of each question within its forbid or allow factor and, thus, distinguish random error inherent to the specific question content. Then, the correlational measure between the forbid abortion factor and the allow abortion factor could be computed, indicating congenericity when the correlation coefficient is 1.0.[3]

The correlational measure needed to test congenericity of forbid and allow questions is most easily obtained by posing one cluster of forbid and an equivalent cluster of allow questions to the same sample of respondents. However, if this were done at one and the same point in time, it would cause memory effects. On the other hand, if forbid and allow were to be posed at different times, it would be impossible to decide whether low correlations should be attributed to attitude change or to the wording effect.

A solution to these interpretation problems is to use a design that allows for a test on congenericity of forbid and allow within the same point in time (so that time effects are excluded) and includes several groups of respondents (so that memory effects can be avoided). To achieve this, Jöreskog's (1971) concept of congenericity has to be generalized to an experimental correlational design involving several groups (van den Bergh, Eiting, and Otter 1988; van den Bergh 1990).

Congenericity means that a forbid factor and an allow factor correlate unity. If forbid and allow are congeneric, it follows that the correlation between two allow factors should equal the correlation between two forbid factors. This inference can be used in constructing a design where the congenericity between forbid and allow questions is tested without disturbing time or memory effects. Note, however, that the reverse is not necessarily true: If the correlations are similar, the congenericity hypothesis for allow and forbid questions cannot be rejected, but this does not imply that congenericity is the case.

A more direct test of congenericity can be obtained by comparing correlations between forbid and allow questions over time. It is not likely, however, that correlation coefficients of 1.0 will be found in that case. Although congenericity refers to a correlation coefficient

between true scores of 1.0, this by no means implies that the respondent's attitude over time is not allowed to change. If we want to test congenericity at two consecutive measurement occasions, we have to relax the definitions somewhat. We define two sets of forbid and allow questions, to be answered on different occasions, congeneric if the correlation between forbid and allow questions (and between allow and forbid questions, for that matter) equals the correlation between allow and allow questions (and between forbid and forbid questions). That is, the wording of the questions does not influence the magnitude of the relation over time.

In this study, the test on congenericity within time will be combined with the comparison over time. The test on congenericity within time will be used to answer the basic question of wording effect researchers: Do forbid and allow measure the same attitude? If forbid and allow are congeneric, then the test on congenericity within time will provide a strong indication for that congenericity, making it unnecessary to compare correlations over time for further evidence. If forbid and allow are noncongeneric, then the test on congenericity within time will provide direct statistical proof of that noncongenericity. In that case, a comparison of correlations between forbid and allow over time to correlations between questions that were formulated similarly at T1 and T2 will provide a touchstone to evaluate the effect size. If the effect of the use of forbid and allow is smaller than the time effect found, then the wording effect might not be important or meaningful for practitioners, in spite of the statistical evidence for the effect provided by the test on congenericity within time. Another advantage of the combination of these two measures of congenericity is that congenericity within time can be tested twice within this design, at both times of measurement.

### *GOALS OF THIS RESEARCH*

The first goal of this research is to test whether forbid and allow questions measure the same attitudes. This question of congenericity was tested using the correlational design outlined in the previous section. This provides a direct test on the different attitudes hypothesis, an explanation of the forbid/allow asymmetry predicting that forbid and allow are noncongeneric for all respondents.

When forbid and allow turn out to be congeneric, further analyses show whether differences in the observed scores for forbid and allow can be attributed to consistent differences between true score or error score variances, or between the response scales. If this is the case, explanations of those differences can, and should, be formulated. This is when the different scales hypothesis will be tested, which predicts that forbid and allow do measure the same trait, and have equivalent error and true score variances, but cause the response scales to be dissimilar interval scales.

In addition, the hypothesis based on the feature-positive effect will be explored, which predicts that forbid and allow are noncongeneric for respondents holding a weak or indifferent attitude toward the issue while being congeneric for people holding a stronger attitude. This exploration will also allow for another check on the congenericity of forbid and allow.

Two different experiments were conducted, the second one being a replication of the first. Both experiments have the same research design and method, but the manipulated questions are about different issues. Converging results from both experiments will support generalizability of the findings.

## *SUBJECTS, INSTRUMENTATION, AND PROCEDURE*

### *INSTRUMENTATION*

For both experiments, a self-administered questionnaire was developed containing several questions formulated with either forbid or allow. To circumvent context effects due to the repetition of forbid/allow questions, and to obscure research goals for the respondents, forbid/allow questions were separated by fillers about related issues.

In each experiment, questions were added to measure attitude strength (cf. Krosnick and Abelson 1992) in order to test the hypothesis based on the feature-positive effect that forbid and allow are noncongeneric for indifferents. The focus was on the intensity of the attitudes ("I have a strong opinion on nature and environmental issues"), the certainty about the expressed opinions ("I was very doubtful on how to answer the questions"), and the importance ("I feel very

involved with . . .") of the issues to the respondents. Attitude strength was not measured per question, only globally per issue (e.g., "nature and environment").

All forbid/allow questions and fillers were answered on a 4-point scale (yes!/yes/no/no!). The responses to questions measuring attitude strength were given on a 5-point scale (from *fully agree* to *fully disagree*).

*Experiment 1*

The subject covered by the questionnaire was nature and environmental issues, since a literature search (in Dutch newspaper databases and policy papers) showed that many recent discussions within this domain are about whether particular phenomena should be forbidden (driving on Sundays, the use of disposable batteries).

The questionnaire contained 36 questions: 5 questions about whether phenomena concerning environmental issues (EI) should be forbidden or allowed (e.g., the use of disposable batteries), 5 questions about whether phenomena concerning nature policy (NP) (e.g., building of new roads through nature areas) should be forbidden (version 1) or allowed (version 2), 16 fillers about related subjects but without the verbs forbid or allow, and 10 questions measuring attitude strength at the end of the questionnaire. Issues were subjects of recent discussion in Dutch public opinion, and all issues were allowed at the time the questions were posed.

*Experiment 2*

The questionnaire developed for the second experiment was generally about public opinion issues concerning ethnic groups. The questionnaire consisted of 5 questions about whether phenomena concerning discrimination (DM) (racist speeches in public, separate classes for Dutch children and children from ethnic groups) should be allowed (version 1) or forbidden (version 2) and of 5 questions about integrational issues (IG) (circumcision for women, nightly call for prayer from the mosque). Again, issues were subjects of discussion in Dutch newspapers, television debates, and policy papers. Contrary to experiment 1, all issues were forbidden at the time the questions were

posed. There were 15 fillers about related issues and 8 questions measuring attitude strength at the end of the questionnaire.[4]

*SUBJECTS AND PROCEDURE*

The experiments were carried out simultaneously. The questionnaires were administered twice, to the same subjects, with a 4-week period between the first and second time of administration. To diminish memory effects as much as possible, there was quite some time in between the first and second administration.

Subjects were students in linguistics and arts. The questionnaire was administered by visiting five different university classes and asking the students in these classes to fill out a questionnaire on environmental issues and ethnic group policies. The students were informed that their responses would assist research on the informativity and feasibility of referendums (which was a hot issue in public opinion around that time). The questionnaire was filled out during class. At the first time of administration, questionnaire versions were randomly distributed within classes. Students were asked to put their names on the empty front page of the questionnaire; it was explained to them that their answers would be stored and treated anonymously. Students' names were needed to trace which ones got which version of the questionnaire, but, to avoid memory effects, subjects were not notified that the questionnaire would be administered again. The total number of students present at the first time of administration was 300, and 297 questionnaires were filled out and returned.

After 4 weeks, the same classes were visited. The students were asked to fill out the questionnaire again to assist research on whether it makes a difference at which time a referendum is held. Questionnaires were handed out on students' names. Half of the students who filled in version 1 the first time received version 1; the other half received version 2. The same applies for the students who filled in version 2 the first time, thus creating a four-group design with about 50 subjects per group. The subjects were not told that there were different questionnaire versions, or that they might have received a different version compared to the second time. After the second administration, the goals of the research were explained to the students. At the second time of administration, some students (98) who cooperated the first

**TABLE 2: Conditions and Number of Respondents per Condition**

| | *Experiment 1* | | | | *Experiment 2* | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *T1* | | *T2* | | *T1* | | *T2* | | |
| *Group* | *EI* | *NP* | *EI* | *NP* | *DM* | *IG* | *DM* | *IG* | n |
| 1 | Forbid | Forbid | Forbid | Forbid | Allow | Allow | Allow | Allow | 47 |
| 2 | Forbid | Forbid | Allow | Allow | Allow | Allow | Forbid | Forbid | 55 |
| 3 | Allow | Allow | Forbid | Forbid | Forbid | Forbid | Allow | Allow | 49 |
| 4 | Allow | Allow | Allow | Allow | Forbid | Forbid | Forbid | Forbid | 44 |

NOTE: EI = environmental issues, NP = nature policies, DM = discrimination, IG = integration.

time did not attend their classes, and a few students (4) refused to cooperate for a second time, resulting in 195 respondents who filled out the questionnaire twice and 136 students who filled out the questionnaire at either T1 or T2. Only subjects who filled out the questionnaire twice were included in the analysis (see Table 2 and Figure 1).

## *DESIGN AND PLANNED ANALYSES*

### *DESIGN*

The main research question is whether forbid and allow are congeneric and, if so, what differences in aspects of the scores cause the observed scores for forbid and allow to differ. The correlational four-group design that was used to answer that question was outlined in previous sections (and in Table 2).

The method of analysis that was used is a structural modeling approach (using LISREL). Models are constructed in which all the variables and relations among them are defined, allowing a statistical test of hypothesized relations between variables (Jöreskog and Sörbom 1986; Bollen 1989). Under the assumption of a multivariate normal distribution, maximum likelihood parameter estimates can be calculated, which are then compared to the observed data. If the difference between the observed data and the computed estimates is small, the model fits the data. The fit of the model is expressed as the likelihood ratio, which is approximately $\chi^2$ distributed. This $\chi^2$ measure, combined with the number of degrees of freedom, can be used as a

measure of relative fit compared to the other models (Jöreskog and Sörbom 1986).

Several models are fitted to the data, and all models are nested within each other. The first model is extremely restrictive, and the last model imposes hardly any restrictions on relations between variables within the data set. The strictest (parallel) model implies that there are no differences in either means, true score variances, or error score variances between forbid and allow questions. The least restrictive (noncongeneric) model describes forbid and allow questions to measure, at least partly, different underlying constructs—in this model, all aspects of the scores for forbid and allow questions differ. The three models in between focus on differences in mean answers (essential tau-equivalent model) between forbid and allow questions, differences in observed variances due to differences in error score variances (tau-equivalent model), and differences in observed variances due to differences in both error score variances and true score variances (congeneric model). By fitting these models, differences between forbid and allow can be analyzed in a stepwise manner.

*PLANNED ANALYSES*

The planned analyses for both experiments consist of three steps. First, descriptive analyses per cluster of questions (EI and NP in experiment 1 and DM and IG in experiment 2) will be conducted, followed by model fitting per experiment over all respondents. Finally, an exploratory analysis focusing on the response behavior of indifferent respondents will be done, providing another check on the congenericity of forbid and allow. Because the analyses do not differ between the two experiments, the method of analysis for both experiments will be worked out simultaneously in this section. All examples will be based on experiment 1, but they can be easily applied to experiment 2.

*PRELIMINARY ANALYSES*

For each cluster of questions (EI and NP) in each version, reliabilities will be computed to check whether the questions indeed measure some attitude related to the issues. If a question in both the forbid and

allow versions does not load on the theorized factor, the item will be excluded from further analysis. Second, to gain some insight into the degree to which the findings of these two experiments are comparable to the asymmetries found in previous forbid/allow experiments, a series of split ballot analyses (per question, per time), will be carried out. Finally, covariance matrices will be computed per experiment, per group, containing the covariances of the observed scores obtained at T1 and T2.[5] These four covariance matrices per experiment will be analyzed simultaneously using LISREL models for multisample analysis (Jöreskog and Sörbom 1986).

*MODEL FITTING*

One after another the models will be analyzed, each nested within another. In each model, one restriction is loosened. The models contain restrictions over relations between clusters of questions (EI and NP), true score variances, and error score variances—within time, and between T1 and T2 (see Figure 1).

Because of the randomization procedure used, there is no reason to assume that characteristics of answers to questions about the same issue that are worded similarly should differ between groups. For all questions about the same issue formulated in the same manner (either forbid or allow) and administered at the same time, two invariant restrictions apply: the regression of observed scores on true scores ($\lambda^{g}_{i,j}$) is the same, and the error score variances ($\theta^{g}_{\varepsilon i,i}$) are the same. A similar restriction holds for the correlations between the factors: The correlations between those clusters have to be the same when factors are administered in the same version, at the same time. Furthermore, the factor variance is restricted to be the same for all groups and for all factors in order to be able to compare the value for $\lambda^{g}_{i,j}$. Without loss of generalizability, the true score variances can be standardized to 1.0 ($\psi^{g}_{j,j}=1$) (van den Bergh and Eiting 1989).

These invariant restrictions over $\lambda$, $\theta_{\varepsilon}$, and $\psi$ are theoretical assumptions about the relations between variables that hold for all models, and will not be tested. Next to these invariant restrictions, variable restrictions are defined, which will be described in the next section.

*VARIABLE RESTRICTIONS*

In this section, the variable restrictions will be described that are used to test hypotheses about the possible differences between answers to forbid and allow questions.

For each experiment, 25 models are specified, all containing restrictions within time and over time. The restrictions within time are the most important restrictions, since it is by these restrictions that the exact differences between forbid/allow questions can be analyzed without a complicating time factor.

The different restrictions within time result in five models that are nested within each other. The restrictions over time are each nested within those five models, resulting in 25 ($5 \times 5$) models. The five models containing restrictions within time will be described. At the end of this section, the restrictions over time will be sketched.

*Restrictions Within Time*

The first model is a parallel model, divided into five submodels (Ia-Ie). These models are all parallel within time, and restrictions over time vary between parallel over time (Ia) to noncongeneric over time (Ie). *Parallelism* within time (Ia-Ie) implies the following restrictions (within and across groups): correlations ($\psi$) within time (within T1 and within T2) between factors (EI and NP) are the same in all groups, independent of the wording of the questions; the regression of observed scores on true scores ($\lambda$) is the same for all equivalent EI questions and for all equivalent NP questions, independent of the wording; error score variances ($\theta_\varepsilon$) are the same for all similar questions, independent of the wording of the questions; and mean scores for all equivalent EI questions and all NP questions are the same, independent of the wording. If one of the parallel models within time fits the data best, there is no difference between forbid and allow questions.[6]

In the *essential tau-equivalent* models within time (IIa-IIe), the restriction is dropped that the mean scores have to be the same for equivalent EI or NP questions that are worded differently. If one of those models fits the data best, means to the questions differ within time, but all other aspects of the score are the same. This means that the wording of the questions causes a difference in scales: the same attitude is

measured with forbid and allow questions, leading to the same true score and error score variance, but the attitude is expressed on a different scale due to the wording of the question. The response scale to forbid and allow questions should be viewed not as a scale on which "yes" means "yes" and "no" means "no" but rather as an interval scale, with different scale midpoints due to the question wording. If one of these essential tau-equivalent models fits best, this would support the different scales hypothesis. Inspection of the means should indicate whether there are consistent patterns in the scale differences for forbid and allow questions.

The *tau-equivalent* models within time (IIIa-IIIe) loosen the restriction that error score variances ($\theta_\varepsilon$) have to be the same for questions that are worded differently at the same point in time. If one of these models fits best, differences in means to forbid and allow questions are caused by differences in the scales and by differences in error score variances.

In *congeneric* models within time (IVa-IVe), all restrictions on question level, apart from the invariant restrictions, are loosened. Only the restrictions over $\psi$ remain: correlations between EI and NP are restricted to be the same over groups. If one of these models fits best, this is an indication that forbid and allow questions measure the same underlying construct: The same attitude is measured with forbid and allow questions within time, but means, error score variances, and true score variances differ. If tau-equivalent or congeneric models fit the data best, and consistent patterns can be found between the differences in true score variances and error score variances, explanations should focus on the differences in reliabilities between forbid and allow questions. Furthermore, differences in true score variances support the different scales hypothesis: the midpoints to the scale differ, as well as the distances between intervals due to the question wording.

In the *noncongeneric* models within time (Va-Ve), all variable restrictions are loosened within time, even the restrictions over ($\psi$). If one of these models fits best, this means that forbid and allow questions measure, at least partly, a different attitude. If this is the case, there is no use in comparing the means or other score characteristics within time. If one of the noncongeneric models fits best, this would support the different attitudes hypothesis.

*Restrictions Over Time*

The restrictions over time follow the same types of models. In *parallel* models, all aspects of the scores are the same over time: Answers to equivalent EI questions have the same mean scores, error score variances, and true score variances, irrespective of wording and independent of time (and the same goes for all NP factors). Furthermore, correlations between EI and NP factors at T1 are similar to those correlations at T2, and correlations between EI (or NP) at T1 and EI (or NP) at T2 are similar between groups. If a parallel model over time fits the data best, there is no attitude change over time. Neither is there any difference in the answers to forbid and allow questions over time: Answers are exactly the same at T1 and T2, irrespective of the similarity or difference in question wording at each time of measurement. In *essential tau-equivalent* models, error score variances do not have to be the same over time for equivalent questions, worded differently or similarly. In *tau-equivalent* models, restrictions over $\lambda$ are loosened. In *congeneric* models, only the correlations between EI and NP are restricted to be the same over time. In *noncongeneric* models, there are no restrictions over time, as restrictions over $\psi$ are loosened as well.

Note that the restrictions over time make no distinction between questions that were worded similarly at T1 and T2 (groups I and IV) and questions that were worded differently at T1 and T2 (groups II and III). This implies that the model that fits best over time says something about the relation between forbid and allow over time, but also about the relation over time between questions that were worded similarly. If a parallel model over time fits the data best, this means two things: There is no difference between forbid and allow and there is no attitude change due to the time lapse. Should a noncongeneric model over time fit the data best, this could mean either that forbid and allow are noncongeneric or that all questions (worded similarly and worded differently) are noncongeneric due to attitude change over time.

## *EXPLORATION: ATTITUDE STRENGTH*

By fitting the 25 models described in the previous sections, the question is answered whether forbid and allow measure the same attitudes, which is a test of the different attitudes hypothesis. The different scales hypothesis is also tested by these models.

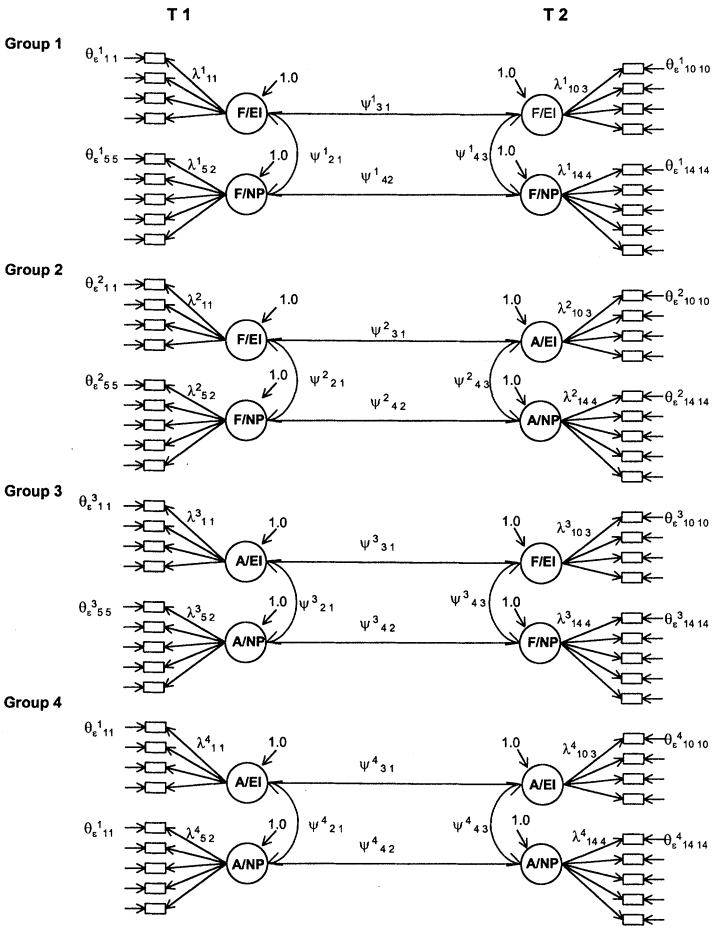


**Figure 1: The Research Design and Some Models**

NOTE: A = allow, F = forbid, EI = environmental issues, NP = nature policies, T1 = the first time of measurement, T2 = the second time of measurement. To illustrate the variant restrictions of the models, model restrictions for the essential tau-equivalent model within time and the non-congeneric model over time are

$$\theta_{\varepsilon ii}^{k} = \theta_{\varepsilon ii}^{k+1} \text{ for } i = 1, 2, \ldots, 18, \text{ and } k = 1, 2, 3,$$

and as the models are nested within each other, the restrictions from less restrictive models apply for this model as well:

tau-equivalent within time: $\lambda_{ij}^{k} = \lambda_{ij}^{k+1}$, for $i = 1, 2, \ldots, 18$, and $j = 1, \ldots, 4$ and $k = 1, 2, 3$;

congeneric within time: $\psi_{21}^{k} = \psi_{21}^{k+1}$ and $\psi_{43}^{k} = \psi_{43}^{k+1}$, for $k = 1, 2, 3$.

An extra test on the (non)congenericity of forbid and allow, in addition to the test on congenericity within time, can be obtained by testing whether the correlation between a forbid factor and a variable $x$ is similar to the correlation between an allow factor and this $x$ variable. If forbid and allow are congeneric, this should logically be the case. In this research, the $x$ variable is the variable of attitude strength.

Furthermore, the hypothesis that the forbid/allow asymmetry is primarily caused by the responding behavior of indifferent respondents, as theorized by Hippler and Schwarz (1986), can be explored by an extra analysis. If forbid and allow turn out to be noncongeneric for the whole sample, it is interesting to test whether this is attributable to the answering behavior of the indifferent respondents. In that case, forbid and allow would be noncongeneric for the subgroup of indifferents while being (at least) congeneric for respondents holding stronger attitudes.

The extra test on congenericity and the exploration of the responding behavior of indifferent respondents can be done by fitting similar models, which will be described below. However, the variable of attitude strength differs somewhat for these two research questions. For both analyses, the answers to the questions measuring attitude strength will be combined into one score for each respondent, per experiment. In each experiment, the attitude strength at T1 and the attitude strength at T2 will be combined into one variable, reflecting the mean attitude strength of a respondent toward both issues (EI and NP). To explore the indifferent respondents hypothesis, the attitude strength variable will be dichotomized on the median, thus creating a dummy variable that isolates respondents with a "weak attitude," whereas the mean scores of the attitude strength variable will be used to test congenericity.[7]

Of the 25 models tested, the attitude strength variable will be introduced to the model that fits best. Three models will be specified: (1) a model that restricts the effect of attitude strength on the factors (EI, NP) to be the same per factor across groups, independent of wording; (2) a less stringent model allowing the effect of attitude strength to differ for factors that are worded differently; and (3) a model that allows for the effect of attitude strength to be estimated freely, differing for factors that are worded differently, and differing for different groups.

These models can be used for the exploration of the effect of attitude strength and for the extra check on congenericity. For the extra test on congenericity, the observed scores for attitude strength are used. If the observed scores on the attitude strength variable are used, and if the first model fits the data best. This provides evidence for the congenericity of forbid and allow. If the second or third model fits best, forbid and allow will be noncongeneric.

For the exploration of the effect of indifferents, the dummy variable of attitude strength is used. If the second or third model fits the data best when using the dummy variable for attitude strength, this would indicate that a low attitude strength causes forbid and allow to be noncongeneric. If the effect of attitude strength differs for factors that are worded differently, or differs for different groups, the correlations between forbid EI and forbid NP cannot logically be similar to the correlations between allow EI and allow NP. Forbid and allow would have to be noncongeneric for the subsample of indifferents. If the first model fits best, this would indicate a lack of support for the hypothesis that forbid and allow measure partly a different attitude for indifferent respondents.

## *RESULTS AND DISCUSSION*

### *PRELIMINARY ANALYSES*

In experiment 1, one question did not load very well on EI in both the forbid and the allow versions and was excluded from further analysis, resulting in an EI factor consisting of four questions and an NP factor consisting of five questions. In experiment 2, both theorized factors consisted of one question that in both versions did not load very well. Both questions were excluded, resulting in a DM factor and an IG factor both consisting of four questions.[8]

For each question posed at T1 or T2, a separate split ballot analysis was done, resulting in 18 split ballot analyses for experiment 1 and 16 split ballot analyses for experiment 2. These analyses indicate that the forbid/allow asymmetry does exist in the present data. Although not for every question, at each time of measurement the key phenomenon itself was replicated (see Table 3 for some examples).

**TABLE 3: Examples of Split Ballot Results for One Question From Each Factor**

| *Factor* | *Question: Do You Think the Government Should Forbid/Allow . . .* | *% Not Forbid* | *% Yes Allow* | $\chi^2$ | p |
|---|---|---|---|---|---|
| NP | . . . military exercises in or nearby nature areas? | T1: 26.5 | T1: 15.4 | 3.53 | .06 |
| | | T2: 19.8 | T2: 23.5 | 0.39 | .53 |
| EI | . . . speeds higher than 90 km/h on highways? | T1: 90.2 | T1: 90.3 | 0.00 | .98 |
| | | T2: 91.7 | T2: 83.8 | 2.77 | .10 |
| DM | . . . exclusion of immigrants for certain job vacancies? | T1: 45.7 | T1: 13.0 | 24.99 | .00 |
| | | T2: 32.3 | T2: 12.8 | 10.48 | .00 |
| IG | . . . marriages to more than one woman for people from polygamous cultures? | T1: 42.9 | T1: 25.5 | 6.49 | .01 |
| | | T2: 39.4 | T2: 27.4 | 3.15 | .08 |

NOTE: EI = environmental issues, NP = nature policies, DM = discrimination, IG = integration.

**TABLE 4: Cumulative Percentages of Not Forbid and Yes Allow Answers per Experiment and per Time**

| *Experiment and Time* | *% Not Forbid (cumulative)* | *% Yes Allow (cumulative)* |
|---|---|---|
| Experiment 1, T1 | 393.1 | 368.8 |
| Experiment 1, T2 | 393.2 | 366.2 |
| Experiment 1, Total | 786.3 | 735.0 |
| Experiment 2, T1 | 288.8 | 184.6 |
| Experiment 2, T2 | 262.4 | 183.0 |
| Experiment 2, Total | 551.2 | 367.6 |

It was explored whether the design used may have confounded the results, whereas it might be the case that having to answer the questions for a second time increased attitude strength and, therefore, caused smaller asymmetry sizes at T2. As can be seen in Table 4, this turned out not to be the case. For both experiments, there is an asymmetry at T1 and at T2. In experiment 2, the asymmetry is smaller at T2 than at T1; in experiment 1, it is the other way around.

### *MODEL FITTING: CONGENERICITY*

The models described in the previous sections were fitted to the resulting data set. The fit of a model can be evaluated by means of a $\chi^2$ testing statistic. The fit of two nested models can be compared by

means of the differences in $\chi^2$ and the differences in the number of degrees of freedom. Because all models are nested within each other, comparison of fit is done by keeping one restriction constant—either over time or within time. In this case, the logical step is to first keep restrictions within time constant, since these are the most important given the research goals. In Table 5, fittings of the congeneric model within time are given for both experiments; restrictions over time are varied in submodels.

The goal within a structural modeling approach like this one is to describe the actual data with a maximum amount of restrictions described with a minimum number of parameters, since this is the model that describes the data maximally economically and extensively. So in comparing the model fit, one looks for the model with the lowest ratio of $\chi^2$ and degrees of freedom.[9]

*Congenericity Over Time*

Comparison of the model fits (with a level of significance of .05) shows that the model that is noncongeneric over time fits the data best, in both experiment 1 and experiment 2 (see Table 5).

Does the fact that the noncongeneric models fit better over time than the other models mean that forbid and allow are noncongeneric? Or does it mean that the attitudes toward the issues changed over time in general, thus affecting the correlational measures between forbid and allow measured at different times? This can be checked by comparing the correlation between questions that were worded similarly at both times (forbid *x* [or allow *x*] at T1 and forbid *x* [or allow *x*] at T2) to questions that were worded differently at T1 and T2. However, this criterion was not specified in the models. So the correlation matrices have to be inspected to find out what the noncongenericity means exactly.

For experiment 1, correlations between factors over time are quite high: Correlations between EI (or NP) at T1 and EI (or NP) at T2 vary between .75 and 1.0. The lowest correlation between NP at T1 and NP at T2 is found in group 2 (forbid at T1, allow at T2), but the lowest correlation between EI at T1 and EI at T2 is found in group 4 (.77), where respondents answered the allow version at both times. So differences in correlations cannot be attributed to wording.

**TABLE 5: Comparison of Models, Restrictions Over Time Varied**

| *Restrictions* | | *Experiment 1* | | | *Experiment 2* | | |
|---|---|---|---|---|---|---|---|
| *Within Time* | *Over Time* | df | $\chi^2$ | *Comparison* | df | $\chi^2$ | *Comparison* |
| IVb Congeneric | Essential-tau | 636 | 999.02 | IVb-IVc: $p = .014$ | 501 | 721.02 | IVb-IVc: $p < .001$ |
| IVc Congeneric | Tau | 591 | 930.71 | IVc-IVd: $p = .524$ | 461 | 644.08 | IVc-IVd: $p = .093$ |
| IVd Congeneric | Congeneric | 573 | 913.72 | IVc-IVe: $p = .011$ | 445 | 620.24 | IVc-IVe: $p = .001$ |
| IVe Congeneric | Noncongeneric | 558 | 876.27 | Best model: IVe | 430 | 583.13 | Best model: IVe |

NOTE: Level of significance used = .05.

This is even more clearly visible in experiment 2, where the correlation between IG at T1 and IG at T2 in group 4 (.64) and the correlation between DM at T1 and DM at T2 in group 1 (.66) is much lower than the correlations obtained in groups 2 and 3 (.72 for DM at T1 and DM at T2 in group 2 to .89 for the correlation between DM at T1 and DM at T2 in group 3), whereas groups 2 and 3 received different versions at T1 and T2, and groups 1 and 4 answered either the forbid or the allow questions at both times.

So in both experiments, the noncongenericity over time is basically caused by instability of the attitudes measured and by instability of the relations between these attitudes (between EI and NP, and between IG and DM), independent of the wording that was used to measure those attitudes. Should forbid and allow turn out to be noncongeneric within time, the size of that wording effect can be evaluated by comparing it to the general time effect found.

*Congenericity Within Time*

Because the influence of time on the attitudes measured was anticipated, the design of this research was provided with a way to measure congenericity of forbid and allow without this disturbing influence. Congenericity within time was operationalized by the restriction that the correlation between the two factors in each experiment (EI and NP in experiment 1; DM and IG in experiment 2) should be similar independent of the wording of the questions in those factors. Of course, the fit of more restrictive models was also tested. These models are a test of the different scales hypothesis by restricting the true score variances to be similar for forbid and allow questions (tau-equivalence) and restricting the true score variances and error score variances to be equivalent (essential tau-equivalence).

In experiment 1, the essential tau-equivalent model fitted the data best (see Table 5): Forbid and allow questions are not only congeneric, do not only measure similar attitudes, they also yield the same true and error score variances. This finding holds for all questions included in this experiment, independent of the specific subissue addressed in each equivalent forbid or allow question. However, the observed scores to equivalent forbid and allow questions differ: The answering scales to forbid and allow questions must be interval scales with

dissimilar midpoints. So the results of experiment 1 support the different scales hypothesis of the forbid/allow asymmetry.

Is it possible to generalize this finding of essential tau-equivalence to all possible forbid and allow questions? Or are the results obtained in experiment 1 partly dependent on the particular issues addressed, or on the fact that all issues addressed in experiment 1 were allowed at the moment the questions were posed? Experiment 2 was an almost exact replication of experiment 1, apart from the fact that the questions were about another subject, and the issues addressed were forbidden at the moment the questions were asked.

Different from experiment 1, in experiment 2, the congeneric model fits the data best (see Table 6).[10] Again, forbid and allow measure similar attitudes, but the error score variances, true score variances, and observed scores differ. Closer inspection of the figures indicates, however, that no systematic patterns in the differences between the error score variances and true score variances to forbid and allow questions can be found.

So, in both experiment 1 and experiment 2, questions worded with forbid or allow are at least congeneric, since in both experiments the correlation between forbid $x$ and forbid $y$ is similar to the correlation between allow $x$ and allow $y$. This is the case within both T1 and T2. Formulated more conceptually, the hypothesis that equivalent questions worded differently (with either forbid or allow) measure the same underlying construct cannot be rejected. This means that the different attitudes hypothesis cannot be accepted. And it also means that it is not necessary to use the measure of congenericity over time as a touchstone to evaluate the effect size.

The results are ambivalent as to whether error score variances and true score variances differ for forbid and allow questions. They are similar in experiment 1 but differ in experiment 2, which may be an indication of issue dependency. But the congenericity in experiment 2 and the essential tau-equivalence in experiment 1 both indicate support for the different scales hypothesis. In experiment 1, only the midpoints of the scales to forbid and allow differ. In experiment 2, the midpoints of the scales differ and the distances between the intervals are dissimilar. Results of experiment 2 indicate another problem too: The reliabilities of forbid and allow questions differ.

**TABLE 6: Comparison of Models, Restrictions Within Time Varied**

| *Restrictions* | | *Experiment 1* | | | *Experiment 2* | | |
|---|---|---|---|---|---|---|---|
| *Within Time* | *Over Time* | df | $\chi^2$ | *Comparison* | df | $\chi^2$ | *Comparison* |
| IIe Essential-tau | Noncongeneric | 621 | 949.48 | IIe-IIIe: $p = .189$ | 486 | 736.55 | IIe-IIIe: $p < .001$ |
| IIIe Tau | Noncongeneric | 567 | 886.56 | IIe-IVe: $p = .178$ | 438 | 621.70 | IIIe-IVe: $p < .001$ |
| IVe Congeneric | Noncongeneric | 558 | 876.27 | IIe-Ve: $p = .164$ | 430 | 583.13 | IVe-Ve: $p = .104$ |
| Ve Noncongeneric | Noncongeneric | 552 | 869.08 | Best model: IIe | 424 | 572.61 | Best model: IVe |

NOTE: Level of significance used = .05.

*EXPLORATION: THE EFFECT OF ATTITUDE STRENGTH*

The explanation based on the feature-positive effect predicts that forbid and allow are noncongeneric for indifferent respondents. Because results of both experiments indicate that forbid and allow are at least congeneric for the whole sample of respondents, this hypothesis is not very plausible. If forbid and allow are noncongeneric for the indifferents, then logically they should be noncongeneric for the sample as a whole, since the correlation coefficients for the subsample would influence the correlation coefficients for the whole sample. So, if this exploration of the indifferents hypothesis indicates noncongenericity for the subsample, this in turn would indicate a power problem of the research reported here.[11]

Three models were fitted to the data in which the restrictions on the influence of attitude strength were varied. These three models were modifications of the model that fitted best for the whole sample: essential tau-equivalent within time and noncongeneric over time for experiment 1 and congeneric within time and noncongeneric over time for experiment 2. Because the noncongeneric restrictions fitted best over time in both experiments, the effect of attitude strength was not restricted to be the same at T1 and T2. Attitude strength was a dummy variable, by which respondents holding a weak attitude toward the issues were isolated.

For both experiments, the first and most restrictive model fitted the data best (see Table 7). This means that the influence of attitude strength is the same for both forbid and allow and the same in every group. Therefore, it can be concluded that forbid and allow are also congeneric for the indifferent respondents. For this group, too, the answers to these differently worded questions reflect the same traits. Furthermore, the results indicate that this research is not plagued by power problems. But the effect of attitude strength was also built in the design as an extra test on the (non)congenericity of forbid and allow. Therefore, the same three models were compared using the observed scores for the attitude strength variable. In this analysis too, the first model fitted the data best: The correlation between forbid questions and attitude strength did not differ from the correlation between allow questions and attitude strength in both experiments. In addition to the test on congenericity within time, this is another strong indication that

**TABLE 7: Effect of Attitude Strength**

| | *Experiment 1* | | | *Experiment 2* | | |
|---|---|---|---|---|---|---|
| *Restriction* | df | $\chi^2$ | *Comparison* | df | $\chi^2$ | *Comparison* |
| Model 1 | 677 | 982.28 | 1-2 $p = .800$ | 482 | 659.14 | 1-2 $p = .418$ |
| Model 2 | 673 | 980.63 | 1-3 $p = .516$ | 478 | 655.23 | 1-3 $p = .121$ |
| Model 3 | 665 | 971.13 | Best model: 1 | 470 | 641.32 | Best model: 1 |

NOTE: Level of significance used = .05.

forbid and allow factors correlate unity, and that forbid and allow questions measure the same trait.

## *CONCLUSIONS AND GENERAL DISCUSSION*

The results of the two experiments reported here show that attitude questions worded with either forbid or allow are at least congeneric: The hypothesis that questions containing either one of those verbs measure different attitudes must be rejected. This is demonstrated by the test on congenericity within time, as well as by the test on congenericity using the extra variable attitude strength. Questions containing either forbid or allow measure the same attitude if random measurement error is filtered out. This means that the different attitudes hypothesis cannot be accepted.

The observed scores for forbid and allow questions differ, however, as the results of split ballot analyses and the bad fit of the parallel models show: The forbid/allow asymmetry is present in both experiments. This means that similar attitudes are expressed differently on the answering scales because of the use of forbid/allow. However, the nature and cause of that difference cannot be allocated unequivocally. The different scales hypothesis is clearly supported in the present data, but it works out differently in both experiments. In experiment 1, only the midpoints of the answering scales differ. In experiment 2, the distances between scale points differ. Furthermore, in experiment 2, reliabilities of forbid and allow questions differ, but no consistent patterns in those differences could be shown.

The findings are ambivalent as to whether forbid and allow questions yield similar error score and true score variances (essential tau-equivalence). The differences in results between experiment 1 and 2 concerning the essential tau-equivalence of forbid and allow could be an indication of a certain degree of issue dependency. It might be that forbid and allow have similar error and true score variances when the issue questioned is not sensitive (nature and environmental issues), whereas the verb pair is "just" congeneric when the questions are about a sensitive subject (such as attitudes toward ethnic groups). More research into this issue dependency seems useful.

But the general conclusion is obvious: The forbid/allow asymmetry does not originate during question interpretation or attitude localization. It springs from the stage in which the opinion is mapped onto the answering options "yes" or "no." The hypothesis offered by Krosnick and Schuman (1988) theorizing the asymmetry to be caused by differences in the way respondents map their answers to the response options due to the use of forbid and allow may well be correct. Forbid and allow questions cause respondents to retrieve or form similar attitudes—perhaps a quite general evaluation, such as "What do you think of abortion?"—that have to be translated from a general negative or positive evaluation into a rather specific "yes or no forbid" or "yes or no allow." The asymmetry results from slight changes in perceptions of the meanings of attitude question response options during this translation process.

An exploration of the effect of attitude strength did not support the hypothesis based on the feature-positive effect, predicting that forbid and allow are noncongeneric for respondents holding a "weak" attitude toward the issues. This exploration did provide extra evidence for the congenericity of forbid and allow.

Attitude strength is an interesting background variable. In this research, no effect of attitude strength on the relations between forbid and allow was found. So, based on an explicit direct measure of respondents' attitude strength, no difference between the answering behavior of respondents holding a weak or a strong attitude could be shown. Self-reports of certainty, intensity, or importance do not seem to be related to forbid/allow answers. Attitude strength could have been operationalized differently, however: Respondents who systematically did not select the extremes of the yes!-no! or the agree-disagree scales could have been defined as indifferents instead of splitting around the median. Also, other direct measures of attitude strength could have been used, or the measures used in this research could have been split up into several factors, each with different effects (Krosnick and Abelson 1992; Waterplas et al. 1988), instead of treating the attitude questions as belonging to one factor.

It seems that more research on the nature of the construct attitude strength is needed in order to come to a firm operationalization of attitude strength and the way it can influence the relations between

questions that are worded differently. Departing from the connotations hypothesis, which states that forbid and allow are both extreme, and which causes the yes-response options to be extreme, one would expect the attitude strength dimension "extremity" to be a more important variable than the ones measured in this research.[12]

The advice to questionnaire designers based on this research is to use either forbid or allow, but to be careful not to interpret the answers in terms of the defined scale points, since the meaning of the scale points and the distance between scale points is dependent on question wording. Both forbid and allow questions are equally valid, but the conclusions based on the answers are not necessarily so.

The policy issues addressed in this research turned out to be non-congeneric over time. This stresses the methodological importance of taking time effects into account when measuring wording effects. The manipulated questions should be administered at the same point in time, or a criterion should be introduced to allocate time effects in the analysis phase. Even a 4-week period is enough to change the attitudes measured, and when this is not taken into account in the experimental design, results of wording effect research will indicate attitude instability instead of wording effects.

A few topics concerning the research reported here deserve some further discussion. First, research following the correlational design described here is complicated enough as it is. It would have been even more complicated if two points answering scales had been used, as was done in previous forbid/allow research, because in that case, polychoric correlations would have had to be computed. In the research reported here, a 4-point scale was used, causing this research not to be an exact replication of previous research. The use of a 4-point scale might have affected the outcomes, since research demonstrates that the number of response alternatives might influence the answers (Cox 1980). On the other hand, respondents still had to choose between "yes" and "no," and analysis of the observed scores does indicate that there was a forbid/allow asymmetry present in the data used in this research.

Second, it would be interesting to repeat this research in a heterogeneous population, for language and art students might be focused enough on linguistics to decrease wording effects. This should preferably be done using a questionnaire by which the theorized

constructs are measured more reliably, thus increasing the fit of each model and making the interpretation of parameter estimates possible and worthwhile.[13]

Finally, as discussed earlier, it would be revealing to relate the extremity of respondents' opinions to the forbid/allow answers they give. For example, one could use a within-subjects design in which respondents answer forbid/allow questions on a yes/no answering scale as well as on a more-points scale so that the meanings of yes/no answers to forbid/allow questions can be compared. This may reveal whether affirmative answers to forbid/allow indeed represent more extreme opinions than negative answers to either question.

The goal of the research reported here, to obtain results that can be generalized across specific question wording variables and specific question topics, seems to have succeeded. The partly conflicting results (essential tau-equivalence in one experiment, congenericity in the other) can now be explained in terms of the general characteristics of the issues addressed instead of focusing on differences between specific questions. The main conclusion to be drawn here is that forbid and allow questions are at least congeneric; that is, the questions measure the same attitude. Answers to forbid and allow questions are expressed differently on the answering scales, however, because of question wording, which supports the different scales hypothesis.

## NOTES

1. A meta-analysis of forbid/allow research, including the data presented here, showed a mean asymmetry size of 14 percent: "Not forbid" was answered 14 percent more than "yes allow." The variance in the asymmetry size turned out to be large (97), and the standard deviation was 9.85 (Holleman 1997).

2. Because the focus is not on individuals but on groups of people, analysis is done over error score and true score *variance* instead of over the actual component of error score or true score.

3. Also, this would, at least for a large part, solve generalization problems, since variation in question wording (apart from the manipulation on forbid/allow) and variation in the nature of the subissues addressed in each question could be achieved.

4. Because of the length of the instruments, space limitations prohibit the provision of full question wordings here. A few examples of questions can be found in Table 3; the complete questionnaires used for this research can be obtained from the author.

5. The observed scores of the forbid questions will be reversed in order to facilitate comparison between forbid and allow answers.

6. By the way, in split ballot designs, the parallel model is compared to all other models, in one go.

7. The measure used for attitude strength is thus rather general and above the level of specific subissues raised in separate forbid/allow questions. Previous research indicates that attitude strength does not have to be measured for each specific act that has to be forbidden or allowed (see Waterplas, Billiet, and Loosveldt [1988], in which the degree of general informedness on public opinion issues is found to be related to the asymmetry size).

8. Preliminary analyses showed that reliabilities of the questions were not very high in either experiment. See Note 10.

9. Not surprisingly, the parallel models turned out not to comply with that demand. The parallel models differ about 1000 $\chi^2$ with a maximum of 70 degrees of freedom compared to less restrictive models (e.g., the parallel-parallel model [Ia] in experiment 1 had 2649.98 $\chi^2$ with 724 degrees of freedom; in experiment 2, it was 1543.15 $\chi^2$ with 579 degrees of freedom), so comparison of fit of all parallel models (Ia-Ie, and all a-models) with the less restrictive models will not be reported in more detail here.

10. As can be seen in Table 4, none of the models fitted the data very well. For model IIe, which fitted best in experiment 1, the goodness of fit index was .665 and the root mean square was .090. For model IVe, which fitted best in experiment 2, the goodness of fit index was .721 and the root mean square was .119. LISREL parameter estimates show reasonably low values for lambda in both experiments. Those weak factor loadings seem to be the main cause for the lack in fit.

So the attitudes in this study have been measured quite unreliably. This is not a problem, however, since the goal of this research was not to develop questionnaires that measure attitudes toward "nature and environmental issues" or "ethnic groups." The goal was to investigate the differences in responses caused by differences in question wording by comparing differences in model fit. Low reliabilities could influence the power of these analyses. Because of the order of the fitted models, and by showing that some models can indeed be rejected, it may be concluded that this lack of power is not very problematic in this study. Because of the bad model fit, however, caution is warranted in interpreting or comparing the parameter estimates.

11. If forbid and allow would have turned out to be noncongeneric for the whole sample, the exploration of the noncongenericity for indifferent respondents would have made more sense. Then the question would have been whether the noncongenericity was caused by the response behavior of indifferents, whereas forbid and allow would have been congeneric for respondents holding stronger attitudes.

12. Indeed, recent experimental research conducted following this hypothesis shows such a relation between extremity of opinions and the meanings of the answers yes and no to forbid/allow questions (Holleman 1999, 2000).

13. Correlations in recent experiments (see Note 12) in a heterogeneous population, in which a yes/no response scale was used instead of a 4-point scale, do turn out to support the findings presented here.

## REFERENCES

Billiet, Jaak. 1989. "Wat te Doen? Beschouwingen over het Nut van Pasklare Voorschriften voor het Ontwerpen van Survey-vragen" [Contemplations on the Use of Ready-Made Prescriptions for Survey Questions]. Pp. 35-52 in *Sociaal Wetenschappelijk Onderzoek met Vragenlijsten. Methoden, Knelpunten, Oplossingen*, edited by Johannes van der Zouwen and Wil Dijkstra. Amsterdam: VU-Uitgeverij.

Bishop, George F., Hans-Juergen Hippler, Norbert Schwarz, and Fritz Strack. 1988. "A Comparison of Response Effects in Self-Administered and Telephone Surveys." Pp. 273-82 in *Telephone Survey Methodology*, edited by R. M. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, and J. Waksberg. New York: Wiley.

Bollen, K. A. 1989. *Structural Equations With Latent Variables*. New York: Wiley.

Cicourel, Aaron V. 1982. "Interviews, Surveys and the Problem of Ecological Validity." *American Sociologist* 17:11-20.

Clark, Herbert H. 1973. "The Language-as-Fixed-Effect-Fallacy: A Critique of Language Statistics in Psychological Research." *Journal of Verbal Learning and Verbal Behaviour* 12:335-59.

Clark, Herbert H., and Michael F. Schober. 1992. "Asking Questions and Influencing Answers, What Is to Be Done?" Pp. 15-48 in *Questions About Questions: Inquiries Into the Cognitive Bases of Surveys*, edited by J. M. Tanur. New York: Russell Sage Foundation.

Cox, Eli P. III. 1980. "The Optimal Number of Response Categories for a Scale: A Review." *Journal of Marketing Research* 17:407-22.

Fazio, Russell H., Steven J. Sherman, and Paul M. Herr. 1982. "The Feature-Positive Effect in the Self-Perception Process: Does Not Doing Matter as Much as Doing?" *Journal of Personality and Social Psychology* 42:404-11.

Hippler, Hans-Juergen, and Norbert Schwarz. 1986. "Not Forbidding Isn't Allowing: The Cognitive Basis of the Forbid/Allow Asymmetry." *Public Opinion Quarterly* 50:87-96.

Holleman, Bregje C. 1997. "Asking Survey Questions: Should It Be Forbidden or Not Be Allowed?" Paper presented at the 7th Annual Meeting of the Society for Text and Discourse, Utrecht, July.

———. 1999. "Wording Effects in Survey Research: Using Meta-Analysis to Explain the Forbid/Allow Asymmetry." *Journal of Quantitative Linguistics* 6:29-40.

———. 2000. "On the Meanings of Yes and No." In *On the Cognitive Mechanisms Underlying Wording Effects: Answering Questions About the Forbid/Allow Asymmetry.*

Jobe, Jared B., and David J. Mingay. 1991. "Cognition and Survey Measurement: History and Overview." *Applied Cognitive Psychology* 5:175-92.

Jöreskog, K. G. 1971. "Statistical Analysis of Sets of Congeneric Tests." *Psychometrika* 36:109-33.

Jöreskog, K. G., and D. Sörbom. 1986. *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood, Instrumental Variables and Least Squares Methods*. Chicago: Scientific Software.

Krosnick, Jon A., and Robert P. Abelson. 1992. "The Case for Measuring Attitude Strength in Surveys." Pp. 177-203 in *Questions About Questions: Inquiries Into the Cognitive Bases of Surveys*, edited by J. M. Tanur. New York: Russell Sage Foundation.

Krosnick, Jon A., and Howard Schuman. 1988. "Attitude Intensity, Importance, and Certainty and Susceptibility to Response Effects." *Journal of Personality and Social Psychology* 54:940-52.

Molenaar, Nico J. 1982. "Response-Effects of 'Formal' Characteristics of Questions." Pp. 49-89 in *Response Behaviour in the Survey-Interview*," edited by Wil Dijkstra and Johannes van der Zouwen. New York: Academic Press.

———. 1986. *Formuleringseffecten in Survey-Interviews, een Non-Experimenteel Onderzoek* [Wording Effects in Survey Interviews, a Non-Experimental Research]. Amsterdam: VU-Uitgeverij.

Nisbett, R. E., and Lee Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice Hall.

Rugg, D. 1941. "Experiments in Wording Questions, II." *Public Opinion Quarterly* 5:91-92.

Schuman, Howard, and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context.* New York: Academic Press.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology.* San Francisco: Jossey-Bass.

Tourangeau, Roger, and Kenneth Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103:299-314.

van den Bergh, Huub. 1990. "On the Construct Validity of Multiple Choice Items for Reading Comprehension." *Applied Psychological Measurement* 14:1-12.

van den Bergh, Huub, and Mindert Eiting. 1989. "A Method of Estimating Rater Reliability." *Journal of Educational Measurement* 26:29-40.

van den Bergh, Huub, Mindert Eiting, and Martha Otter. 1988. "Differentiële Effecten van Vraagvorm bij Aardrijkskunde en Natuurkunde Examens" [Differential Effects of Item Format in Geography and Science Exams]. *Tijdschrift voor Onderwijsonderzoek* 13:270-84.

Waterplas, Lina, Jaak Billiet, and Geert Loosveldt. 1988. "De Verbieden Versus Niet Toelaten Asymmetry: Een Stabiel Formuleringseffect in Survey-Onderzoek?" [The Forbid/Allow Asymmetry: A Stable Wording Effect in Survey Research?]. *Mens en Maatschappij* 63:399-415.

*Bregje Holleman is an assistant professor in the Department of Social Science Methodology at Free University Amsterdam. Her research aims at explaining wording effects by means of theoretical analysis, experimental research, and meta-analytic techniques.*